Stochastic Gradient Descent Learning for Text Classification

Natural Language Processing

Prof. Amir Jafari

December 12, 2022

I. Introduction

About 80% of all information in the world is unstructured, with text being the most common form of unstructured data (MonkeyLearn). In addition, the rapid growth of online information, has made text classification increasingly important as a method for organizing text data. Using text classifiers, entities can structure text such as emails, social media post, chatbots and more. Even more, text classification can be used to determine the sentiment of text, intent of a message, or detect the language in text. There are several methods can be used to perform text classifications, with most hailing from rule-based systems, machine learning-based, or hybrid systems. However, this paper explores the use of machine learning-based systems, specifically the Logistic Regression Algorithm and the Support vector Machine (SVM). The ability of these models to perform classification can be improved through learning processes. Specifically, the Stochastic Gradient Descent (SGD) learning.

II. Dataset Description

"Information on more than 180,000 Terrorist Attacks

The Global Terrorism Database (GTD) is an open-source database including information on terrorist attacks around the world from 1970 through 2017. The GTD includes systematic data on domestic as well as international terrorist incidents that have occurred during this time period and now includes more than 180,000 attacks. The database is maintained by researchers at the National Consortium for the Study of Terrorism and Responses to Terrorism (START), headquartered at the University of Maryland."

https://www.kaggle.com/datasets/START-UMD/gtd?select=globalterrorismdb_0718dist.csv

Content

Geography: Worldwide

Time period: 1970-2017, *except 1993*

Unit of analysis: Attack

Feature: Summary of attack

Variables: >100 variables on location, tactics, perpetrators, targets, and outcomes

Sources: Unclassified media articles (Note: Please interpret changes over time with caution. Global patterns are driven by diverse trends in particular regions, and data collection is influenced by fluctuations in access to media coverage over both time and place.)

Definition of terrorism:

"The threatened or actual use of illegal force and violence by a non-state actor to attain a political, economic, religious, or social goal through fear, coercion, or intimidation."

III. Description of Models

Logistic regression is one of the more common supervised models in machine learning and is often used for classification and predictive analytics. It is considered a discriminative model, because it attempts to distinguish between classes. The Logistic regression model works by estimating the probability of an event occurring. Since the outcome is a probability, the outcome/dependent variable is mapped between 0 and 1.

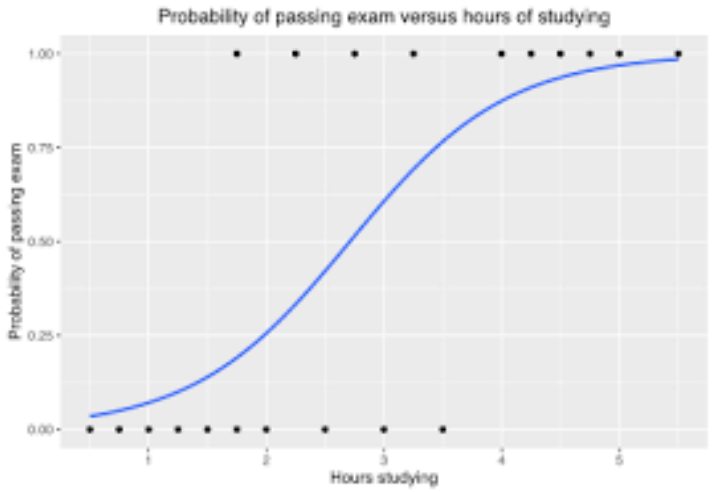


Figure 1. Example of Logistic Regression

SVM is a robust classifier that is good at maximizing the predictive accuracy of a model without overfitting the training data (IBM). It is not used as much today but it still performs well at analyzing large data and it has many applications such as facial image recognition, text

mining. SVM works by mapping data to a high-dimensional feature space so data points can be categorized even when they don't seem separable. The line that separates these classes is referred to as the optimal hyperplane, because it is the most optimal line or plane that maximizes the distance of the margin between the classes

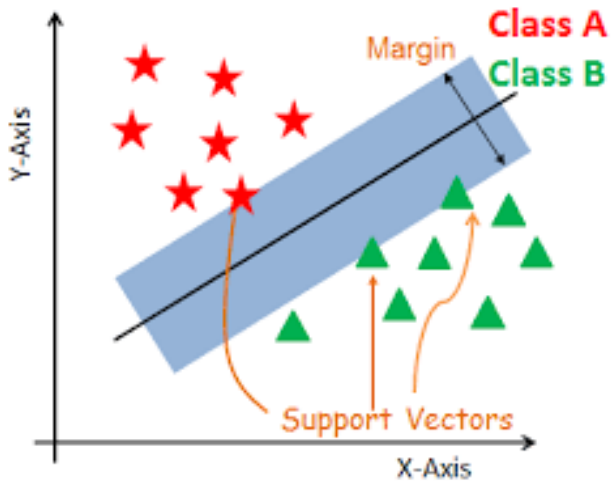
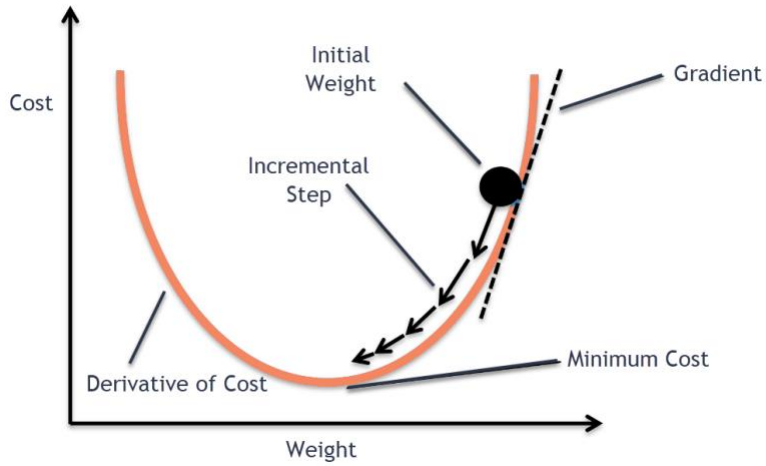


Figure 2. Example of SVM

Lastly, the optimizer that is used in this experiment to optimize these models is the Stochastic Gradient Descent (SGD). The SGD is a subset of Gradient Descent. It is an optimization algorithm that aims to find the coefficients of a model that minimizes the cost or error. It performs different coefficient values and the cost function estimates their cost through the predicted result of each sample taken. This above-mentioned process is done iteratively, with a sample taken at random to go through the process, hence the name stochastic.



IV. Experimental Setup

Collecting and Preprocessing Data

The GTD raw data had been collected by START, hosted by University of Maryland. The data initially had 180,000 observations but after removing null values was reduced to roughly ~110,000 observations. And the relevant columns summary and attack type were kept. After, cleaning and processing was done on the summary column to remove special characters and stop words.

Feature Extraction

Features were extracted by transforming the summary text to numerical features using TF-IDF vectorizer (specified ngram_range of (1,3)) which does:

1. Converting text to numeric feature vectors by using 3-gram representations
2. Weighing the count features by using TF-IDF transformation

TF assigns weight based on its occurrence in the document, while IDF is the number of documents containing that word/token. After SMOTE was applied on the minority classes, to match the number of observations for the class with the highest count(attack_type = bomb).

Before fitting the gridsearch on the data and applying hyperparameter tuning, the feature was scaled using MaxAbsScaler().

Classification Design

The experiment aims to compare the performance of Logistic Regression and SVM In three different case using sklearn

1. Using default parameters w/out SGD learning
2. SGD learning
3. Applying SGD learning with hyperparameter turning (gridsearch)
    - Parameters being tuned loss and learning rate

Validation and Evaluation

- Accuracy: To calculate the percentage of samples in the test set that the classifier correctly labeled, measured by calculating TP+TN/TP+FP+FN+TN [2]
- Precision: Indicates how many incidents were relevant, and measured by TP/ (TP+FP)
- Recall: Indicates how many of the relevant items were identified, and measured by TP/ (TP+FN)
- F1 combines the precision and recall to give a single score. It is the harmonic mean of the precision and recall and calculated by (2 × Precision × Recall)/ (Precision Recall) (Diab et. al 2018)

Results

Classifier	**Performance Metrics**

	Accuracy	Precision	Recall	F1-Score
Log	**87.3**	**0.796**	**0.590**	**0.644**
SVM	**85.81**	**0.793**	**0.498**	**0.542**
Log + SGD	**83.92**	**0.526**	**0.427**	**0.453**
SVM + SGD	**85.83**	**0.819**	**0.503**	**0.549**
SVM + SGD + Tuning	**88**	**0.768**	**0.621**	**0.667**

{'alpha': 0.0001, 'loss': 'hinge', 'penalty': 'l2'}

---

^Best parameters

This experiment evaluated the approach on using unseen testing data on a Logistic and SVM modeling comparing the use of SGD learning and without SGD Learning, and optimizing the SGD learning algorithm with gridsearch. The results were interesting and unexpected. It was thought that SGD incorporated models would outperform the based models without. But the opposite happened, the models without SGD learning performed better or just as good as the models without. The log outperforms the SVM, Log+SGD, SVM+SGD model by a larger margin in all metrics with an accuracy of 87.4. It is the second-best performing model. However, the best performing model in terms of accuracy is the SVM+SGD+tuning with an accuracy of 88. And outperforms the log model is all metrics but precision.

The results of the grid search that found the SVM+SGD+tuning to be best performing model with a learning rate of 0.0001 and penalty of “l2.”

Conclusion

This research proposed Stochastic Gradient Descent learning and Gridsearch hyperparameter tuning in an attempt to optimize text classification on terrorisms attacks. Logistic Regression and SVM specifically, were used to classify terrorist attacks. Regardless of surveyed literature, this experiment cannot be used to determine whether SGD improves performance of Logistic Regression and SVM. We can also conclude that hyperparameter tuning has the potential to optimize the accuracy of the selected classifiers. This experiment was still able to confirm that Logistic Regression and SVM can be used to achieve a fair accurate score. More work is still needed to determine why the addition of the SGD learning decreased the performance of the two classifiers. This was an unexpected result being that current literature has shown that is can optimize models

Future Direction

Being that there is a variety of other classifiers, it would be interesting to compare their classification abilities against one another. In addition, there are other optimizers that can be used to build robust models.

Reference:

Bassey, Patricia. "Logistic Regression vs Support Vector Machines (SVM)." *Medium*, Axum Labs, 19 Sept. 2019, https://medium.com/axum-labs/logistic-regression-vs-support-vector-machines-svm-c335610a3d16.

Diab, S. (2019). Optimizing Stochastic Gradient Descent in Text Classification Based on Fine-Tuning Hyper-Parameters Approach. A Case Study on Automatic Classification of Global Terrorist Attacks. *ArXiv, abs/1902.06542*.

"How SVM Works." *IBM*, https://www.ibm.com/docs/en/spss-modeler/saas?topic=models-how-svm-works.

"Introduction to SGD Classifier - Michael Fuchs Python." *MFuchs*, 11 Nov. 2019, https://michael-fuchs-python.netlify.app/2019/11/11/introduction-to-sgd-classifier/.

"Text Classification: What It Is and Why It Matters." *MonkeyLearn*, https://monkeylearn.com/text-classification/#:~:text=This%20is%20where%20text%20classification,fast%20and%20cost%2Deffective%20way.